Real Dataset Analysis

This analysis compares distinct approaches for modelling competing risks using a high-dimensional dataset on non-muscle-invasive bladder cancer from Dyrskjøt et al. (2004). This dataset includes gene expression (1,381 variables), clinical information, records of death (for bladder cancer, other causes, and not registered), and the time to the event (progression-free survival). In total, the dataset contains 404 observations.

We will compare the performance of this multinomial model against a penalized binary Cox proportional hazards model. In this common approach, a separate model is fitted for each event type, treating all other competing events as censored. This dataset has previously been analyzed by Tapak et al. (2015) using cause-specific Cox models with LASSO, elastic net, SCAD, and SICA penalizations to identify prognostic gene signatures.

1 Preprocess data

First, the two gene expression files were loaded, combined, and transposed to create a dataset where each row is a patient and each column is a gene probe. Next, the clinical data was loaded and cleaned by handling missing values and formatting key variables for survival analysis, including the survival `time` and a categorical `event` for competing outcomes. A minor adjustment was made, converting any survival times of 0 to 0.001 to prevent computational errors.

Finally, the gene expression and clinical datasets were merged by a unique `sample_id`. The resulting dataset was then filtere to include only the patient cohort used in the original study (Dyrskjøt et al. 2004) and to remove samples with missing data. As a result, the dataset kept 301 out of the 404 original observations.

```r
# Process microarray data ---
bladder_fpd1 <- read_delim(here::here("paper/data/GSE5479_Final_processed_data_1.txt"))
bladder_fpd2 <- read_delim(here::here("paper/data/GSE5479_Final_processed_data_2.txt"))

bladder_fpd <- t(cbind(bladder_fpd1[, -1], bladder_fpd2[, -1])) %>%
  data.frame() %>%
  tibble() %>%
  bind_cols(tibble(sample_id = c(names(bladder_fpd1[, -1]),
                   names(bladder_fpd2[, -1])))) %>%
  set_names(c(bladder_fpd1$probe, "sample_id")) %>%
  clean_names() %>%
  select(sample_id, everything())

# Process clinical information
```

```r
bladder_hd <- read_xls(here::here("paper/data/6517200/10780432ccr062940-sup-supplemental_
  clean_names() %>%
  mutate(across(everything(), \(x)case_when(x == "-" ~ NA,
                            T ~ x))) %>%
  transmute(
    sample_id = case_match(sample_id,
                   "1082-1" ~ "1082-1_DK",
                   "20421_S (91?)" ~ "20421_S",
                   .default = sample_id),
    country,
    # Survival time and event
    event = progression_0_no_progression_1_progression_to_t1_2_progression_to_t2,
    time = as.numeric(progression_free_survival),
    # time = as.numeric(follow_up_total),
    # Adjust time = 0
    time = ifelse(time == 0, 0.001, time),
    # Clinical variables
    age = as.numeric(age),
    female = if_else(str_trim(sex) == "F", 1, 0),
    progression = factor(
      progression_0_no_progression_1_progression_to_t1_2_progression_to_t2),
    clinicalrisk = clinical_risk_1_high_risk_0_low_risk,
    followup = follow_up_months_from_tumor_to_last_visit_to_the_clinic_or_to_cystectomy,
    ## Reclassification of NA based on paper
    treatment = case_when(is.na(bcg_mmc_treatment) ~ "No treatment",
                  T ~ bcg_mmc_treatment),
    cystectomy = cystectomy,
    grade = reevaluated_who_grade_no_reevaluation,
    stage = reevaluated_pathological_disease_stage_no_reevaluation,
    # Identify samples used in the original paper's model training/validation
    progmodel = as.numeric(!is.na(samples_used_for_training_progression_classifier) | !is
 )


# Create complete bd
bladder_comp <- bladder_hd %>%
  mutate(sample_id = case_when(sample_id == "692-1" ~ paste0(sample_id,
                                        "_",
                                        country),
                     T ~ sample_id)) %>%
  select(-country) %>%
  full_join(bladder_fpd %>%
          #' Creating individual ID's for repeated ids.
          #' It assumes that the samples follow the same order as in
          #' supplementary file.
          mutate(sample_id = case_when(sample_id == "692-1...144" ~ "692-1_F",
                             sample_id == "692-1...145" ~ "692-1_DK",
                             T ~ sample_id)),
        by = join_by(sample_id)) %>%
  filter(progmodel == 1,
      !is.na(age),
      !is.na(female)) %>%
  select(-sample_id, -clinicalrisk, -cystectomy,
      -progmodel, -followup, -progression)
```

2 EDA

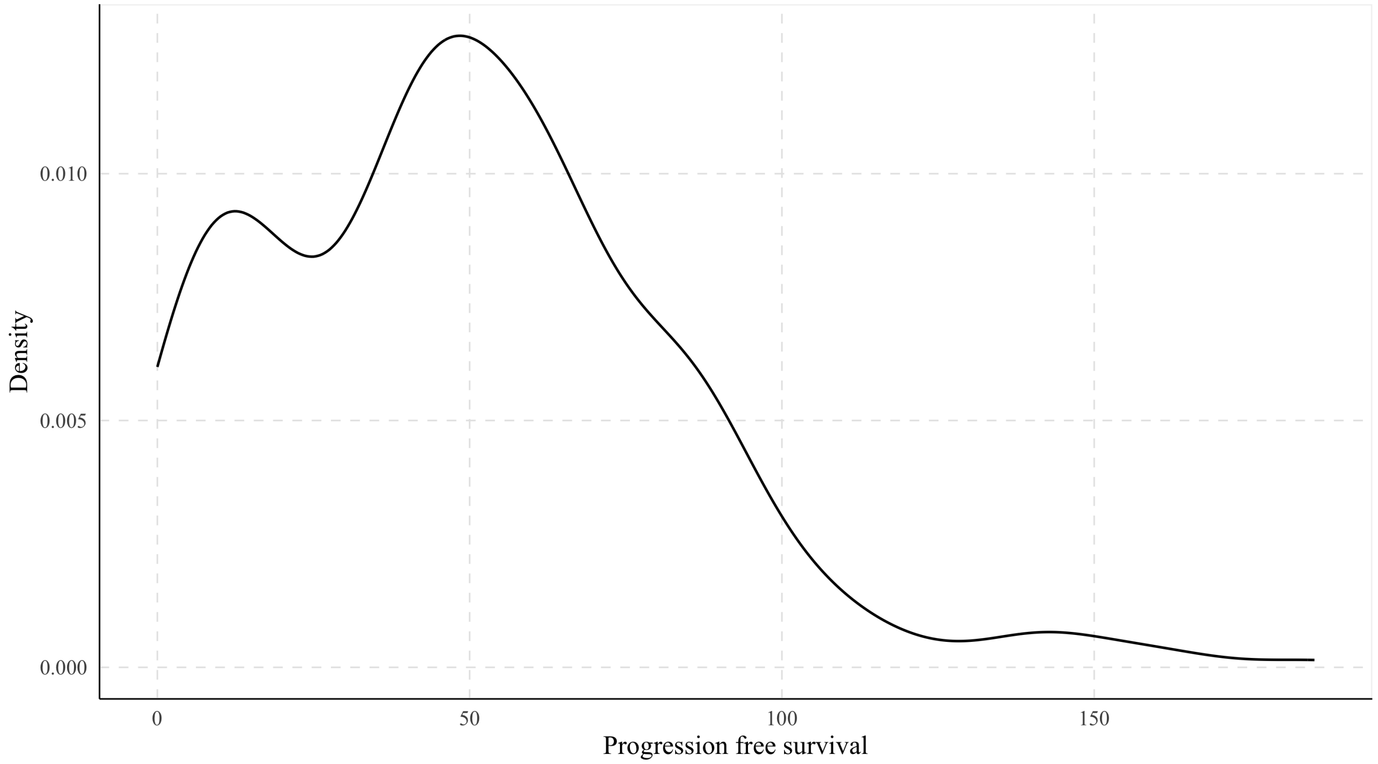
The majority of patients (227) were censored, 20 patients died from bladder cancer and 54 patients died from other causes.

```
#" Count events
bladder_comp %>%
  count(event) %>%
  adorn_totals()
```

```
#>  event   n
#>      0 227
#>      1  20
#>      2  54
#>  Total 301
```

The distribution of progression-free survival time is multi-modal, with a notable peak around 50 months and a long right tail. The average progression-free survival time across the cohort was 49.5 months. The population time plot visualizes the decrease in the at-risk population over the study's duration, with points indicating when progression events occurred.

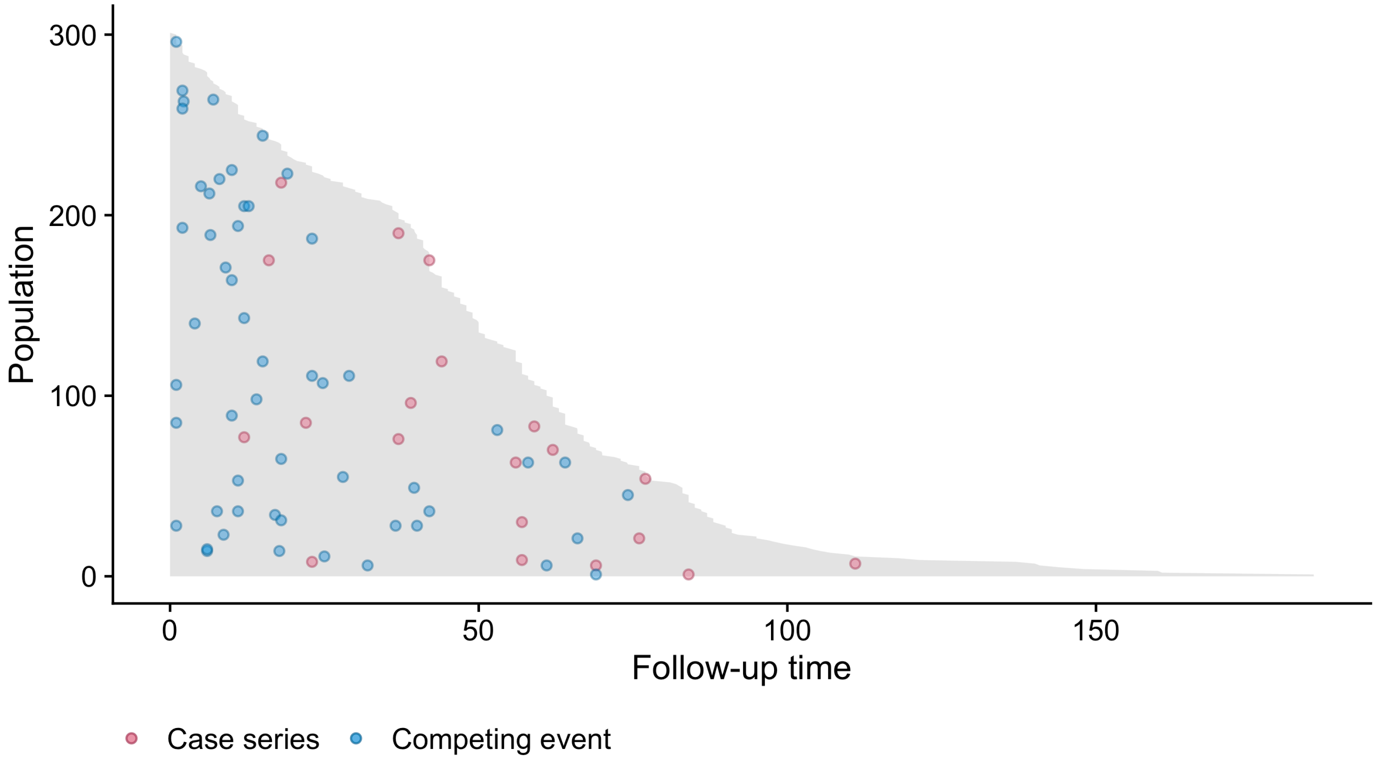
```
# Time
## Time dist
bladder_comp %>%
  ggplot(aes(x = time)) +
  geom_density() +
  labs(x = "Progression free survival",
     y = "Density")
```



```r
## Mean time
mean(bladder_comp$event)
```

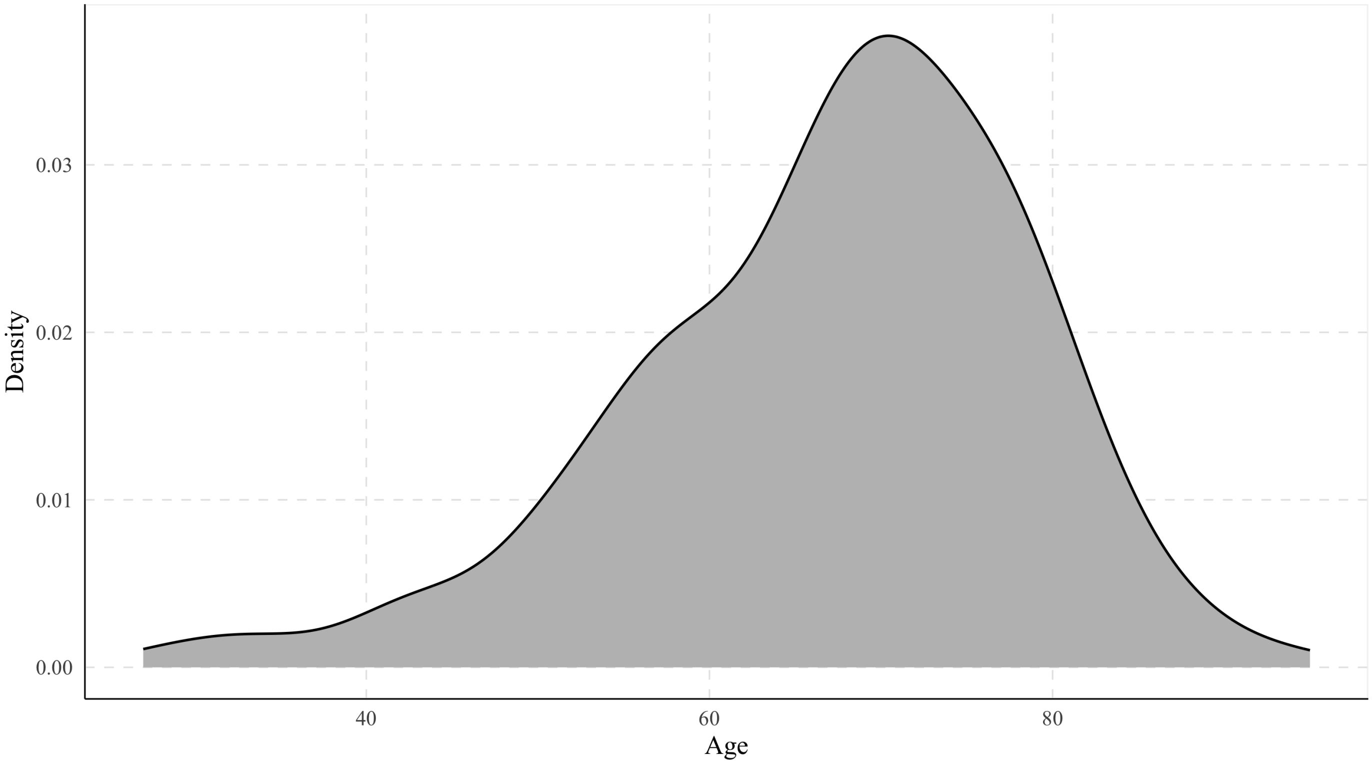
```
#> [1] 0.4252492
```

```r
## Population time plot
plot(popTime(bladder_comp, "time", "event"),
     add.competing.event = TRUE,
    comprisk = TRUE)
```

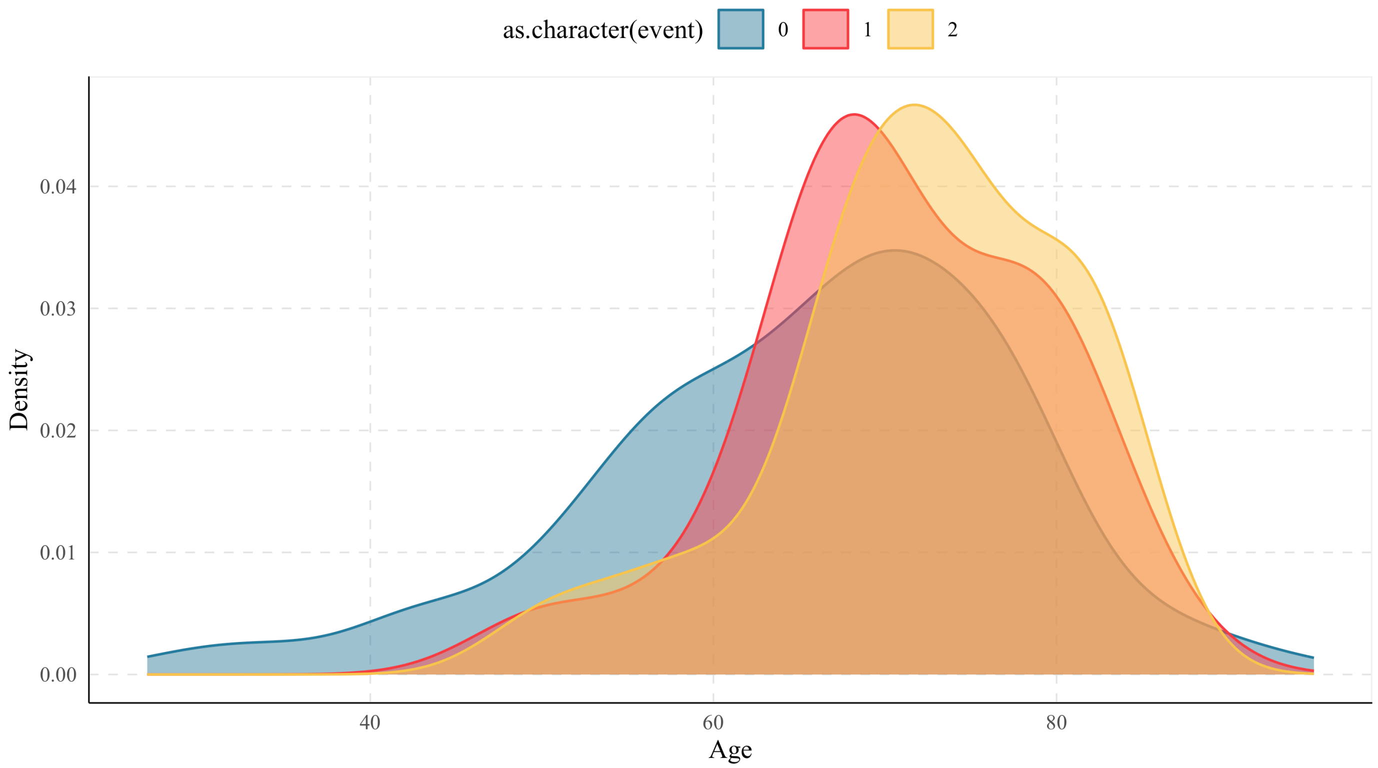


The patient population is predominantly male (241 males vs. 60 females), with an age distribution that peaks around 70 years.

```
# Clinical variables
## Age dist
bladder_comp %>%
  ggplot(aes(x = age,)) +
  geom_density(fill = "grey") +
  labs(x = "Age",
     y = "Density")
```



```
bladder_comp %>%
  ggplot(aes(x = age, colour = as.character(event),
         fill = as.character(event))) +
  geom_density(alpha = 0.5) +
  labs(x = "Age",
     y = "Density")
```



```r
## Categorical var. levels
bladder_comp %>%
  select(female:stage) %>%
  mutate(across(everything(), as.character)) %>%
  pivot_longer(everything(),
          names_to = "Var",
          values_to = "Category") %>%
  group_split(Var, .keep = T) %>%
  map(~count(.,Var, Category) %>%
  adorn_totals())
```

```
#> [[1]]
#>    Var Category   n
#> female        0 241
#> female        1  60
#>  Total        - 301
#>
#> [[2]]
#>   Var Category   n
#> grade     HIGH 175
#> grade    HIGH*   2
#> grade      LOW  80
#> grade  LOW OBS   1
#> grade     LOW*  11
#> grade   PUNLMP  32
#> Total        - 301
#>
#> [[3]]
#>   Var Category   n
#>  stage     T1b   1
#>  stage     pT1  23
#>  stage    pT1*   2
#>  stage    pT1a  54
#>  stage    pT1b  47
#>  stage     pTa 160
#>  stage pTa obs   2
#>  stage    pTa*  11
#>  stage    pTis   1
#>  Total       - 301
#>
#> [[4]]
#>       Var     Category   n
#>  treatment          BCG  73
#>  treatment    BCG - MMC   4
#>  treatment          MMC   5
#>  treatment No treatment 219
#>      Total            - 301
```

Analysis of the clinical categories indicates that most patients had high-grade tumors and were at the pTa stage. A significant portion of the cohort (219 patients) did not receive intravesical BCG or MMC treatment.

Based on these categories, grade, stage, and treatment were recategorized as done in Ke, Bandyopadhyay, and Sarkar (2023). The PUNLMP (Papillary Urothelial Neoplasm of Low Malignant Potential) was classified as a low-grade. For the stage, all derived subcomments were removed, leaving only the pTa, pTis, and T1 categories. In comparison to Ke, Bandyopadhyay, and Sarkar (2023), they removed the observation with pTis stage. Lastly, treatment was classified as non versus either BCG or MMC. Another relevant difference from Ke, Bandyopadhyay, and Sarkar (2023) is that they categorized age; however, we do not have a clear reason to create arbitrary groups.

It is worth noting that the baseline categories were set to low for grade, pTa for stage, and none for treatment. This is especially relevant as the coefficients depend on the baseline group. Additionally, this can explain a different selection compared to previous studies using case-specific Cox models. For example, Tapak et al. (2015) used these models, but the preprocessing for the clinical variables and baseline categories was not found.

```
# Adjust categories
bladder_comp_adj <- bladder_comp %>%
  mutate(grade = str_remove(grade, "\\*| OBS"),
      grade = case_when(str_detect(grade, "PUNLMP") ~ "LOW",
                        T ~ grade),
      grade = fct_relevel(grade, "LOW"),
      stage = case_when(str_detect(stage, "pTa") ~ "pTa",
                        str_detect(stage, "T1|pTis") ~ "T1",
                        T ~ stage),
      stage = fct_relevel(stage, "pTa"),
      # No treatment as base
      treatment = ifelse(str_detect(treatment, "BCG|MMC"), 1,
                   0))

bladder_comp_adj %>%
  select(female:stage) %>%
  mutate(across(everything(), as.character)) %>%
  pivot_longer(everything(),
          names_to = "Var",
          values_to = "Category") %>%
  group_split(Var, .keep = T) %>%
  map(~count(.,Var, Category) %>%
  adorn_totals())
```

```
#> [[1]]
#>     Var Category   n
#>  female        0 241
#>  female        1  60
#>   Total        - 301
#>
#> [[2]]
#>    Var Category   n
#>  grade     HIGH 177
#>  grade      LOW 124
#>  Total        - 301
#>
#> [[3]]
#>    Var Category   n
#>  stage       T1 128
#>  stage      pTa 173
#>  Total        - 301
#>
#> [[4]]
#>        Var Category   n
#>  treatment        0 219
#>  treatment        1  82
#>      Total        - 301
```

```
## Categorical var. levels

bladder_comp_adj <- model.matrix(~ .,
                    data = bladder_comp_adj)[,-1] %>%
  as_tibble()

saveRDS(bladder_comp_adj,
     here("paper", "data", "bladder_comp_adj.rds"))
```

References

Dyrskjøt, Lars, Mogens Kruhøffer, Thomas Thykjaer, Niels Marcussen, Jens L. Jensen, Klaus Møller, and Torben F. Ørntoft. 2004. "Gene Expression in the Urinary Bladder: A Common Carcinoma in Situ Gene Expression Signature Exists Disregarding Histopathological Classification." *Cancer Research* 64 (11): 4040–48. https://doi.org/10.1158/0008-5472.CAN-03-3620.

Ke, Chenlu, Dipankar Bandyopadhyay, and Devanand Sarkar. 2023. "Gene Screening for Prognosis of Non-Muscle-Invasive Bladder Carcinoma Under Competing Risks Endpoints." *Cancers* 15 (2): 379. https://doi.org/10.3390/cancers15020379.

Tapak, Leili, Massoud Saidijam, Majid Sadeghifar, Jalal Poorolajal, and Hossein Mahjub. 2015. "Competing Risks Data Analysis with High-Dimensional Covariates: An Application in Bladder Cancer." *Genomics, Proteomics & Bioinformatics* 13 (3): 169–76. https://doi.org/10.1016/j.gpb.2015.04.001.